Parallel Corpus of Croatian-Italian Administrative Texts

Marija Brkic Bakaric
Department of Informatics
University of Rijeka
Croatia

mbrkic@uniri.hr

Ivana Lalli Pacelat
Faculty of Interdisciplinary, Italian and Cultural Studies
Juraj Dobrila University of Pula
Croatia
ilalli@unipu.hr

Abstract

Parallel corpora constitute a unique resource for providing assistance to human translators. The selection and preparation of the parallel corpora also conditions the quality of the resulting MT engine. Since Croatian is a national language and Italian is officially recognized as a minority language in seven cities and twelve municipalities of Istria County, a large amount of parallel texts is produced on a daily basis. However, there have been no attempts in using these texts for compiling a parallel corpus. A domain-specific sentence-aligned parallel Croatian-Italian corpus of administrative texts would be of high value in creating different language tools and resources. The aim of this paper is, therefore, to explore the value of parallel documents which are publicly available mostly in pdf format and to investigate the use of automatically-built dictionaries in corpus compilation. The effects that a document format and, consequently sentence splitting, and the dictionary input have on the sentence alignment process are manually evaluated.

1 Introduction

Parallel corpora constitute a unique resource, not only for the development of machine translation (MT) systems, but also for providing assistance to human translators. They have been used to develop computer-assisted translation (CAT) tools and resources for human translators, such as translation memories (TM), terminology management tools and resources, bilingual concordances, and translator oriented word processors (cf. McEnery and Xiao, 2007; Kenning, 2010, Somers, 2001). The selection and preparation of the parallel corpora also conditions the quality of the resulting MT engine, since both dominant approaches to MT, statistical machine translation (SMT) and neural machine translation (NMT), rely on high quality parallel corpora.

In bilingual or multilingual areas in which the equal status of two or more languages is officially recognized, a large amount of parallel texts is produced on a daily basis. Due to the officiality of the minority languages and the official nature of the texts and of the context of language use, having a precise and uniform terminology as well as developed translation/language technologies that facilitate the whole translation process is of high importance (Trosterud, 2002). In order to improve the quality of translation, to reduce the time and the cost of the translation, and to preserve the official bilingualism and multilingualism, a number of actions have been initiated over the years in different bilingual and multilingual countries, regions or organizations. The full insight into the tools and resources necessary for facilitating and supporting the multilingual text production is given by the European Commission (Steinberger et al., 2014; European Commission, 2016). Supports have been given to the minority language engineering with a focus on MT development (e.g. for the Basque-Spanish language pair (Alegria et al., 2005) and for the Catalan-Spanish language pair (Arranz et al., 2006)), on terminology (e.g. for Welsh (Jones and Prys, 2006), for Italian, German and Ladin (Streiter et al., 2004)), and on parallel corpus building (e.g. the Trilingual Allegra-Corpus of German, Italian and Romansh (Scherrer and Cartoni, 2012), the Hansard French-English corpus and The United Nations Parallel Corpus v1.0 (Ziemski et al, 2016)).

Unfortunately, this is not the case of Istria County in Croatia, where existing parallel texts

Temnikova, I, Orăsan, C., Corpas Pastor, G., and Mitkov, R. (eds.) (2019) *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, September 5 - 6, pages 11–18.
https://doi.org/10.26615/issn.2683-0078.2019_002

have not been used so far for compiling a parallel corpus needed for MT and other human language technology (HLT) applications.

According to the Statute of the Istrian County (Art. 6, 21, 22, 23, and 24/2009), the Croatian and the Italian language are in equal official use in institutions of the County and of the official bilingual cities and municipalities. The Italian language is officially recognized as a minority language in seven cities and twelve municipalities in Istria County. Due to the equal status of Italian and Croatian, legal and administrative documents have to be published in both languages. The texts are usually written in Croatian and then translated into Italian.

The analysis of the current translation practice and terminology use shows that there is a need to develop translations tools and language resources which would enable a more efficient and faster translation process and ensure the usage of precise and unambiguous Italian terminology in Croatia.

Although parallel corpora for both Croatian and Italian exist, they are mostly in combination with English, as emphasized by Tadić et al. (2012) for Croatian and Calzolari et al. (2012) for Italian. There are also few parallel corpora including both languages of interest, Croatian and Italian – the OPUS2 parallel corpus (Tiedemann, 2012), the EUR-Lex Corpus (Baisa et al., 2016), the Eur-Lex judgments corpus (Baisa et al., 2016), the DGT-Translation Memory (Steinberger et al., 2012), the EAC-TM, the InterCorp (Čermák and Rosen, 2012), the Bulgarian-X language Parallel Corpus (Koeva et al., 2012), etc. These corpora, although few of them belong to the public administration domain, cannot fully satisfy the needs of the local translators and cannot be considered high quality corpora for facilitating the development of translation technology due to the specific bilingual terminology. Since Italian, which is a national language in Italy, has a minority language status in Croatia, differences and particularities of the two legal systems should be taken into account and a consistent and comprehensive Italian terminology adapted to the Croatian legal system should be prepared and used accordingly. The availability of parallel texts abundant in the respective terminology makes the goal of preparing a high quality domain-specific parallel corpus achievable.

Therefore, the aim of this work is to create a domain-specific sentence-aligned parallel Croatian-Italian corpus of administrative texts, which would be valuable in the Istrian case for the creation of different language tools and resources. Sentence alignment is the task of mapping the sentences of two given parallel corpora which are known to be translations of each other. Since the problem of correct sentence alignment is additionally burdened by erroneous sentence splitting (Biçici, 2007), in this paper we explore the value of parallel documents which are publicly available mostly in pdf format.

The research conducted in this paper can be divided into two parts. The first part is related to the preparation of the parallel documents and the second to sentence alignment. Since dictionary input affects sentence alignment, one line of this research explores the difference between sentence alignment without a dictionary input and sentence alignment with a dictionary input. Although both methods rely on the dictionary usage, the first makes use of the dictionary compiled from the same parallel corpus based on the sentence length information, while the latter makes use of the dictionary compiled from another corpus, similar in nature, which is already sentence-aligned.

Related work is presented in section 2. Section 3 deals with the corpus preparation and is divided into corpus and dictionary descriptions, and the description of automatic sentence alignment procedure. Evaluation of the sentence alignment approaches and of the dictionary compiled from the corpus which is not sentence-aligned is given in section 4. A short conclusion along with the directions for future work is given in the last section of the paper.

2 Related Work

The aim of this work is similar to the work in Soares and Krallinger (2019) and Doğru et al. (2018). Soares and Krallinger (2019) build two bilingual and one trilingual corpus for MT purposes and then build NMT models and evaluate translations according to the BLEU score. They conduct evaluation of randomly selected 100 sentences per corpus and mark them as “correct”, “partial”, or “no alignment”. Although in this work we use the labels as in Aker et al. (2014), their meaning is the same. Doğru et al. (2018) gather and prepare medical parallel corpora for the purpose of MT training. The authors report the automatic and semi-automatic methods they use for creating domain-specific (medical) custom translation memories as well as bilingual terminology lists, which include

web-crawling, document alignment in CAT tools and term extraction.

Etchegoyhen et al. (2018) acknowledge that domain-specific resources are usually scarce. However, it is widely accepted that MT works better with domain-specific parallel corpora (Doğru et al., 2018). Evaluation of the benefits of domain adaptation for MT, on three separate domains and language pairs, with varying degrees of domain specificity and amounts of available training data is presented by Etchegoyhen et al. (2018). Doğru et al. (2018) believe that concentrating on the parallel corpora selection, collection and preparation processes is equally important and may have a positive impact on the MT system quality and post-editing.

The first part of the research is similar to the one in Aker et al. (2014). There are three main approaches to the problem of sentence alignment: length-based, dictionary-based, and similarity-based (Varga et al., 2007). In this work we focus on the dictionary-based method and investigate two approaches. The authors in Aker et al. (2014) additionally propose and apply three cleaning methods to the noisy dictionary created by GIZA++. In a method-by-method comparison the transliteration method performs the best, however, the combination of the methods proves to have the highest precision. In this paper we do not apply any dictionary cleaning methods. Our focus is drawn to spurious line breaks introduced by pdf to plain text conversion since, due to the structure of the administrative documents, a simple deletion of these line breaks would badly affect the sentence splitting procedure.

3 Preparation of the Corpus

Since Italian is officially recognized as a minority language in seven cities and twelve municipalities in Istria County, legal and administrative documents of the County and of these official bilingual cities and municipalities have to be published in both Italian and Croatian.

Parallel documents are collected from the Web using a semi-supervised approach. A manual examination of the web sites reveals that suitable parallel documents exist on only four web sites (Istria county[1], Novigrad[2], Pula[3], Umag[4]). We decide to restrict ourselves to the official gazettes as these are published the most frequently of all the bilingual content available. We exclude those sites that publish two-column bilingual pdf files in which the text in Croatian is in one column, and the text in Italian in another column or those that just partly translate the content.

Due to the diversity of web page languages and formats, the python library *Beautiful Soup* and the command *wget* are used for extracting URLs and automatically fetching documents. The identified web sites containing potential parallel documents are first manually inspected and then different types of content within these websites are recognized. Finally, the URLs of official gazette editions are acquired and the respective documents fetched. The alignment on a document-level is performed based on the analyzed and manually detected naming conventions.

Since the downloaded files are mostly in pdf format, the conversion to plain text format is performed. Some basic pre-processing is also conducted, such as removing redundant spaces and empty lines. Please note that the documents contain a lot of numerical data which might give exaggerated perception of the size.

As evident from Table 1, less than half of the Croatian (hr) official gazette editions are available in Italian (it).

Subcorpus	**# of hr docs**	**# of it docs**	**# of aligned parallel docs**
Istria	660	429	251
Novigrad	126	106	65
Pula	65	57	37
Umag	286	71	70

Table 1: Number of documents per corpus.

[1] https://www.istra-istria.hr/index.php?id=8
https://www.istra-istria.hr/index.php?id=486

[2] http://www.novigrad.hr/hr/administracija/dokumenti/category/sluzbene_novine
http://www.novigrad.hr/it/administracija/dokumenti/category/sluzbene_novine

[3] http://www.pula.hr/hr/opci-podaci/sluzbene-novine/
http://www.pula.hr/it/dati-generali/bollettino-ufficiale/

[4] http://www.umag.hr/hr/gradska-uprava/sluzbene-novine-grada-umaga?syear=
http://www.umag.hr/it/gradska-uprava/sluzbene-novine-grada-umaga?syear=

3.1 Sentence-Aligned Parallel Corpus

The software *hunalign* (Varga et al., 2007) is used for sentence alignment. The tool can be run by providing a dictionary but also without one. If no dictionary is provided, *hunalign* resorts to Gale and Church algorithm which is based on the notion that character lengths of source and target sentences are correlated. A dictionary is built based on such alignment, and then the second iteration of the algorithm does the realignment by combining sentence length information with the dictionary. If a dictionary is provided as input, the first step is skipped.

Input files contain Croatian and Italian corpora, both segmented into sentences (one sentence per line) and into tokens (delimited by space characters). We use a version of the tokenizer provided with the *moses* toolkit[5] to which we add the abbreviation list for Croatian[6]. The output contains the aligned sentences (one aligned sentence per line). The entire process of building the sentence-aligned Croatian-Italian corpus of Istria county and cities is shown in Figure 1.

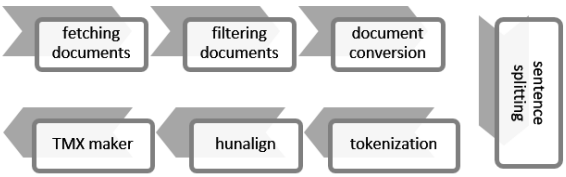

Figure 1: Building the Croatian-Italian corpus.

Since the structure of the public administration documents is such that they contain a wealth of long titles, subtitles, tabular data, lists, references, etc., which often span over multiple lines, the conversion from pdf to plain text format results in many spurious line breaks. We determine that removing these line breaks badly affects sentence splitting, i.e. titles and subtitles stay merged, data from multiple cells stay merged, list items often stay merged, etc. If there are no sentences with appropriate sentence markers in-between, a multi-line text might even end up as a single line. Therefore, we keep the splits introduced by the format conversion and can thus talk about segment splitting rather than sentence splitting.

The descriptions of the four subcorpora of which our corpus consists are given in Table 2. In parallel, we select only those documents that are originally in doc format and perform the steps shown in Figure 1.

Subcorpus	# lines	# tokens	
		Croatian	Italian
Istria	1.2M	3.2M	3.4M
Novigrad	378K	1.2M	1.7M
Pula	318K	858K	1.0M
Umag	638K	1.8M	2.3M

Table 2: Number of lines and tokens per corpus.

3.2 Dictionary

We download the freely available DGT's translation memory (DGT-TM) (Steinberger et al., 2012). We use it for producing a sentence-aligned parallel Italian-Croatian corpus of the European Union's legislative documents (Acquis Communautaire). The corpus statistics is presented in Table 3.

# of	Italian	Croatian
sentences	284 864	284 864
words	5 501 552	4 669 480
characters	38 281 881	34 233 328

Table 3: Description of the sentence-aligned DGT corpus used for automatic dictionary building.

The translation memory mostly consists of the Acquis Communautaire documents. Due to some pre-processing, the contents of the original documents might have somewhat changed. We process 1267 tmx documents and extract 284 864 Italian-Croatian sentence pairs.

A bilingual dictionary is automatically generated using the GIZA++ tool (Och and Ney, 2003), similarly to Aker et al. (2014). One of the major drawbacks of the tool, as the authors in Aker et al. (2014) point out, is the difficulty in using it for technically non-sophisticated users. In addition, the parallel corpus needs to be pre-processed prior to running the tool. Since every source language word is treated as a possible translation of every target language word, the dictionaries created by GIZA++ contain a lot of noise. Words with high translation probabilities may still be wrong. However, we do not perform any filtering at this point of time and only pre-process the dictionary to put it in a format suitable for *hunalign*.

[5] http://www.statmt.org/moses/
[6] https://github.com/clarinsi/reldi-tokeniser

The entire process of creating the dictionary to be used as input for the alignment process is shown in Figure 2. The dictionary contains 793 803 entries.

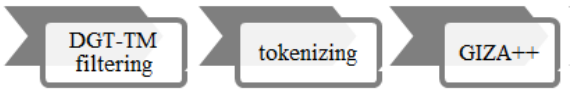

Figure 2: Dictionary creation pipeline.

4 Evaluation

4.1 Evaluation of Sentence Alignments

We conduct manual evaluation of the aligned pairs. The assessment is done by two different evaluators. We randomly select 100 aligned pairs in such a way that all four sub-corpora are represented proportionally to their size and that translation units starting with digits or one-word units are discarded. Aligned pairs are labelled as equivalent (label *equiv.*) if the target segment is an acceptable translation of the source segment, as containment (label *cont.*) if the entire source segment is acceptably translated by a proper sub-part of the target language segment, and none of the above (label *none*) if neither of the first two options applies (Aker et al., 2014). The results of evaluations of sentence alignments on the whole corpus, of sentence alignments on Novigrad subcorpus which is originally in doc format, and of sentence alignments based on the DGT dictionary on Novigrad subcorpus are presented in Table 4, Table 5, and Table 6, respectively. The first row gives sums of evaluations per category, while the second row shows only cases for which there is agreement. The precision is calculated by dividing the number of equal evaluations with the total number of evaluations considered. The interrater agreement is from substantial to almost perfect with the Cohen's kappa scores 66%, 92%, and 73%, respectively (Cohen, 1960). The interpretation of the scores is taken over from Landis and Koch (1977).

As evident from Table 4 and Table 5, the precision is affected by the line breaks introduced with pdf-to-txt conversion, which cannot be solved straightforwardly without affecting the sentence splitting procedure. The precision increases greatly if we consider only documents in doc format. However, the difference in sentence alignment performed by *hunalign* without the DGT-based dictionary and with the DGT-based dictionary is not pronounced. Therefore, it can be concluded that DGT-based dictionary adds no value to the sentence alignment process. This might prove different if we were to use some kind of dictionary filtering.

In order to have more reliable precision results, the evaluation might be amended with an arbitration phase, where a third annotator would judge the cases where the first two annotators disagree. Such approach is taken by Mihalcea and Pedersen (2003) in the evaluation of word alignment.

	Equiv.	**Cont.**	**None**	**Precision**
Sum of evaluations	92	56	52	46%
Evaluations in agreement	40	18	20	78%

Table 4: Evaluation of global sentence alignment without dictionary input.

	Equiv.	**Cont.**	**None**	**Precision**
Sum of evaluations	172	11	17	86%
Evaluations in agreement	85	5	8	98%

Table 5: Evaluation of sentence alignment on word processing documents without dictionary input.

	Equiv.	**Cont.**	**None**	**Precision**
Sum of evaluations	170	6	24	85%
Evaluations in agreement	82	1	10	93%

Table 6: Evaluation of DGT-dictionary-based sentence alignment on word processing documents.

4.2 Evaluation of Dictionary

We also perform a manual evaluation of the automatically built Istrian-based dictionary by randomly selecting 100 different highest probability dictionary entries. We follow the same evaluation methodology as in the previous subsection.

Table 7 presents manual evaluation results. The Cohen's kappa score is almost 69% meaning that there is substantial agreement between evaluators according to the interpretation given by Landis and Koch (1977).

	Equiv.	**Cont.**	**None**	**Precision**
Sum of evaluations	115	32	53	57.5%
Evaluations in agreement	52	8	19	79%

Table 7: Dictionary evaluation.

5 Conclusion and Future Work

The aim of this work is to create a domain-specific sentence-aligned parallel Croatian-Italian corpus. Such resource could be used for training an MT system, automatic terminology extraction, domain adaptation, etc. However, it seems there is a need to correct/validate alignment pairs when working with public administration documents converted from pdf. This would greatly enhance the quality of parallel corpus.

Based on the results of this research, in our future work we plan to extend our corpus and experiment with different methods for compiling or cleaning the dictionary, e.g. neural network-based word alignment, active learning, etc.

Creating such a valuable resource would enable us to train MT systems or to perform domain-adaptation on generic Croatian-Italian MT systems and thus facilitate the work of our public administration. For example, manually revised domain-specific terms extracted from such a resource would enable applying a domain adaptation technique available for SMT which adds phrasal term translations as favored translation options using the XMLmarkup functionality.

Acknowledgments

This work has been fully supported by the University of Rijeka under the grant number 17.14.2.2.01 and the bilateral Croatian-Slovenian project (2018-2019) of the Ministry of Science and Education.

References

Ahmet Aker, Monica Lestari Paramita, and Robert Gaizauskas. 2014. Bilingual Dictionaries for All EU Languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Languages Resources Association (ELRA), Reykjavik, Iceland, pages 2839–2845. http://www.lrec-conf.org/proceedings/lrec2014/pdf/803_Paper.pdf.

Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Aingeru Mayor, et al. 2005. An Open Architecture for Transfer-based Machine Translation between Spanish and Basque. In *Proceedings of the MT Summit X Workshop. Workshop on Open-Source Machine Translation.* Asia-Pacific Association for Machine Translation (AAMT), pages 7–14. https://pdfs.semanticscholar.org/d346/7010dd32f2f317f66cdc0bb532fcb045a97b.pdf.

Victoria Arranz, Elisabet Comelles, and David Farwell. 2006. Speech-to-Speech Translation for Catalan. In Isabella Ties, editor, *Proceedings of the Lesser Used Languages and Computer Linguistics Conference (LULCL 2005)*. Accademia Europea Bolzano, Bolzano.

Vít Baisa, Jan Michelfeit, Marek Medveď, and Miloš Jakubíček. 2016. European Union Language Resources in Sketch Engine. In *The Proceedings of tenth International Conference on Language Resources and Evaluation (LREC 16)*. European Language Resources Association (ELRA). Portorož, Slovenia, pages 2799–2803. https://www.aclweb.org/anthology/L16-1445.

Ergun Biçici. 2007. Local Context Selection for Aligning Sentences in Parallel Corpora. In Boicho Kokinov, Daniel C. Richardson, Thomas R. Roth-Berghofer, editors, *Modeling and Using Context*. CONTEXT 2007. Lecture Notes in Computer Science, volume 4635 (82–93). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74255-5_7.

Nicoletta Calzolari, Bernardo Magnini, Claudia Soria, and Manuela Speranza. 2012. *The Italian language in the digital age / La lingua italiana nell'era digitale.* Berlin: Springer Verlag. http://www.meta-net.eu/whitepapers/e-book/italian.pdf.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20(1), pages 37–46.

European Commission. 2016. *Translation tools and workflow.* Luxemburg: Publication Office of the

European Union. https://doi.org/DOI: 10.2782/703257.

František Čermák and Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3), 411–427. https://doi.org/10.1075/ijcl.17.3.05cer.

Gökhan Doğru, Adrià Martín, and Anna Aguilar-amat. 2018. Parallel Corpora Preparation for Machine Translation of Low-Resource Languages: Turkish to English Cardiology Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Paris, France, pages 12–15. http://lrec-conf.org/workshops/lrec2018/W3/pdf/5_W3.pdf.

Thierry Etchegoyhen, Anna Fern, Andoni Azpeitia, and Eva Mart. 2018. Evaluating Domain Adaptation for Machine Translation Across Scenarios. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, pages 6–15. https://www.aclweb.org/anthology/L18-1002.

Dewi B. Jones and Delyth Prys. 2006. The Welsh National Online Terminology Database. In *Proceedings of the Lesser Used Languages and Computer Linguistics Conference (LULCL 2005)*. Accademia Europea Bolzano, Bolzano, Italy, pages 149–169.

Marie-Madeleine Kenning. 2010. What are parallel and comparable corpora and how can we use them? In Anne O'Keeffe and Michael McCarthy, editors, *The Routledge handbook of corpus linguistics*. Routledge, London, pages 487–500. https://doi.org/10.4324/9780203856949.ch35.

Svetla Koeva, Ivelina Stoyanova, Rositsa Dekova, Borislav Rizov, and Angel Genov. 2012. Bulgarian X-language Parallel Corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, pages 2480–2486. http://www.lrec-conf.org/proceedings/lrec2012/pdf/587_Paper.pdf.

Richard Landis and Gary Koch. 1997. The Measurement of Observer Agreement for Categorical Data Data for Categorical of Observer Agreement. *Biometrics*. 33(1), pages 159–174. https://doi.org/10.2307/2529310.

Tony McEnery and Richard Xiao. 2007. Parallel and comparable corpora: What are they up to? In Gunilla M. Andermann and Margaret Rogers, editors, *Incorporating Corpora: Translation and the Linguist* (Translating Europe). Multilingual Matters, Clevedon, pages 1–13.

Rada Mihalcea and Ted Pedersen. 2003. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*. Association for Computational Linguistics Stroudsburg, PA, USA, pages 1–10. https://www.aclweb.org/anthology/W03-0301.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*. 29(1), pages 19–51. https://doi.org/10.1162/089120103321337421.

Yves Scherrer and Bruno Cartoni. 2012. The Trilingual ALLEGRA Corpus: Presentation and Possible Use for Lexicon Induction. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), Istanbul, Turkey, pages 2890–2896. http://www.lrec-conf.org/proceedings/lrec2012/pdf/685_Paper.pdf.

Felipe Soares and Martin Krallinger. 2019. BVS Corpus : A Multilingual Parallel Corpus of Biomedical Scientific Texts. CoRR,arXiv:1905.01712 [cs.CL], pages 1–8. https://arxiv.org/abs/1905.01712.

Harold Somers. 2001. Bilingual Parallel Corpora and Language Engineering. In *Proceedings.of the Workshop on Language Engineering for South-Asian Languages.* pages 1–16. http://www.emille.lancs.ac.uk/lesal/somers.pdf.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A Freely Available Translation Memory in 22 Languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), Istanbul, Turkey, pages 454–459. http://www.lrec-conf.org/proceedings/lrec2012/pdf/814_Paper.pdf.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the European Union's highly multilingual parallel corpora. *Lang. Resour. Eval.* 48(4), pages 679–707. https://doi.org/10.1007/s10579-014-9277-0.

Oliver Streiter, Mathias Stuflesser, and Isabella Ties. 2004. CLE, an Aligned, Tri-lingual Ladin-Italian-German Corpus. Corpus Design and Interface. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. European Language Resources Association (ELRA), Lisbona, Portugal, pages 84–87. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.616.3181&rep=rep1&type=pdf.

Marko Tadić, Dunja Brozović-Rončević, and Amir Kapetanović. 2012. *The Croatian language in the digital age / Hrvatski jezik u digitalnom dobu.* Hei-

delberg: Springer Verlag. http://www.meta-net.eu/whitepapers/e-book/croatian.pdf.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, pages 2214–2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Trond Trosterud. 2002. Parallel corpora as tools for investigating and developing minority languages. In Lars Borin, editor, *Parallel corpora, Parallel worlds. Language and Computers*. Studies in practical linguistics no 43. Rodopi, Amsterdam, pages 111–122. https://doi.org/10.1163/9789004334298_007.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel Corpora for Medium Density Languages. In *Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005*, Current Issues in Linguistic Theory 292, pages 247–258. https://doi.org/10.1075/cilt.292.32var.

Michał Ziemski, Marcin Junczys-Dowmun, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *The Proceedings of tenth International Conference on Language Resources and Evaluation (LREC 16)*. European Language Resources Association (ELRA). Portorož, Slovenia, pages 3530–3534. https://www.aclweb.org/anthology/L16-1561.